LOST IN SPACE: GEOLOCATION IN EVENT DATA

SOPHIE J. LEE, HOWARD LIU, AND MICHAEL D. WARD

ABSTRACT. Extracting the "correct" location information from text data, i.e., determining the place of event, has long been a goal for automated text processing. To approximate human-like coding schema, we introduce a supervised machine learning algorithm that classifies each location word to be either correct or incorrect. We use news articles collected from around the world (Integrated Crisis Early Warning System [ICEWS] data and Open Event Data Alliance [OEDA] data) to test our algorithm that consists of two stages. In the feature selection stage, we extract contextual information from texts, namely, the N-gram patterns for location words, the frequency of mention, and the context of the sentences containing location words. In the classification stage, we use three classifiers to estimate the model parameters in the training set and then to predict whether a location word in the test set news articles is the place of the event. The validation results show that our algorithm improves the accuracy rate of the current geolocation methods of dictionary approach by as much as 25%.

Keyword: Natural Language Processing, Information Extraction, Text Analysis, Supervised Machine Learning, Geolocation, Event Data

1. INTRODUCTION

Many quantitative studies of conflict rely on event data. Recently, these studies have also retreated from the country-year framework and have focused on disaggregating the event flows both in terms of space and time. Disaggregating temporality—even to the daily level—is a straightforward task. But figuring out precisely *where* an event actually occurred is a difficult and uncertain task that has been perplexing for most contemporary event data efforts (Boschee *et al.* 2015), PHOENIX (OEDA 2016), SCAD (Salehyan 2015), ACLED (Raleigh *et al.* 2010).

At the same time there is widespread interest in disentangling investigations from the country-year framework. The country-year—as an observational framework—has a longstanding tradition in political science. Indeed two of the three most cited articles in the *American Political Science Review* focus on the country-year (Beck & Katz 1995; Fearon & Laitin 2003). As recently as 2009, collections have focused on disaggregation of the country-year in conflict studies (Cederman & Gleditsch 2009). More broadly, many focus on hierarchical approaches that simultaneously include subnational, national, and even international aspects. Efforts at the World Bank

Date: November 16, 2016.
We appreciate the comments of Kyle Beardsley, Andrew Hall, Jan Kleinnijenhuis, Sayan Mukherjee, Jonathan Nagler, Molly Roberts, Colin Rundel, David Siegel, Brandon Stewart, Joshua Tucker, members of Wardlab at Duke University, and the SMaPP lab at New York University. This research was partially supported by the National Science Foundation Award 1259266.
Corresponding Authors: ophie J. Lee and Howard Liu.

and other international organizations (Frank & Martinez-Vazquez 2014; Easton *et al.* 2011) have emphasized this deeper dive into the political and economic landscape. Much of this deeper dive is coming from organizations and governments in terms of their reporting on demographic, economic, financial, and health data that are subnational.

Most data in the conflict realm comes from non-offical sources. For many that means some form of data collected from historical and journalistic sources. This need is often filled by event data, which are typically collected on a daily basis, and can be aggregated temporally to the level required by the analysis. Event data can also be aggregated to the geographical region that is appropriate. Given the increasing demands for event data, the scientific community has recently devoted significant efforts to automate the data collection process. Having humans read and code a large set of archive documents sometimes limits reproducibility, and hence hinders scientific research. It is also expensive and limits the currency of the data. Further, ensuring inter-coder reliability is challenging, especially over global events that span decades.

Several efforts utilize machine-coding to collect event information and determine event features automatically. Projects such as the Integrated Crisis Early Warning System (ICEWS) (Boschee *et al.* 2015; Lautenschlager *et al.* 2015; O'Brien 2010) and the Open Source Event Data Alliance 2016 are two prominent examples. A good overview on event data in political science is found in Schrodt & Yonamine (2013). These automated event data allow researchers to observe and extract information on politically relevant events around the world in near real-time.

Despite the apparent advantages of automated data collection, the machine-coding ontologies for event data require further research (Grimmer & Stewart 2013; Lucas *et al.* 2015). NSF currently sponsors a multidisciplinary project to look into formulating the generation of event data. [1] At the same time, IARPA is reportedly looking to fund an event data challenge that could lead to new ways of collecting and analyzing event data. This attention by funding agencies illustrates that not all issues in this research domain are yet resolved. Outstanding issues include machine translation of texts in foreign languages (wherein great progress is being made in both Chinese and Arabic), duplicate reports from multiple sources, and the relatively low accuracy in determining the event location (D'Orazio *et al.* 2014; Schrodt 2015). While all of these are important, in this article, we focus on the sole issue of geolocation in event data.

For human coders, locating events by reading a news article may be time-intensive, but straightforward. This is not the case for machine-coding: many news articles contain multiple location names, such as the location of the journalist writing the story, the birthplace of a person being interviewed, or the place of a similar event that occurred several decades ago; at times, human names are identical to geographic names; and location names are transliterated into English in a variety of potentially confusing ways. All these sources of *noise* in the data increase the difficulty in automatically locating events. A good algorithm should read texts like human coders and code only the correct event locations.

We treat the geolocation task as a classification problem where each location word is predicted to be correct or incorrect. With the goal of developing an algorithm to discern correct event locations automatically, we extract the contextual information of location words (the N-gram patterns for location words, the frequency of mention, and the context of the sentences

[1] Modernizing Political Event Data for Big Data Social Science Research, Patrick T. Brandt (PI, EPPS), Vito D'Orazio (Senior Personnel, EPPS), Jennifer S. Holmes (Co-PI, EPPS), Latifur R. Khan (Co-PI, ECS), Vincent Ng (Co-PI, ECS), National Science Foundation, RIDIR, $1,497,358, September 2015—August 2018.

containing location words) from the training set. We check the accuracy against a hand-coded set of ground truth data on locations. To do so, estimated parameters from the training set were used to predict the event locations in the test set (for which we also know the ground truth), using three classification methods: artificial neural networks with back propagation, support vector machine (SVM), and random forests. Our supervised machine learning language model codes locations correctly at an accuracy rate close to 90% when the texts contain a single correct location per article.[2]

Our approach is fully automated, but does require some hand coding of a small number of stories to contextualize the coders for specific countries. While the process described in this article is generalizable, we selected data from ICEWS (Boschee *et al.* 2015) and OEDA.[3] We began with an investigation of 250 protests in China (CAMEO code 14: protest), but to anneal the generalizability of our approach, we added (and coded) 250 violent events in the Democratic Republic of the DRC (CAMEO code 19: fight) drawn from the ICEWS data, and 250 violent events drawn from the PHEONIX data on Syria (OEDA, CAMEO code 19: fight).[4] Each of these cases presents difficult problems for automated geolocation.

2. Automated Geolocation

The task of determining event locations involves three steps, each non-trivial. In the first step, known as *named entity recognition* in Computer Science and Computational Linguistics (D'Orazio *et al.* 2014; Cardie & Wilkerson 2008; Guerini *et al.* 2008; Arguello *et al.* 2008; Nadeau & Sekine 2007), all location names are identified and extracted from an appropriately preprocessed text. This step is a prerequisite for the other steps because to determine the location of an event in a news article, capturing the exhaustive list of location names is required. Next comes entity disambiguation/resolution, which involves identifying the actual location of the recognized name string(Cucerzan 2007; D'Orazio *et al.* 2014; Bunescu & Pasca 2006). Once this is accomplished, it is possible to extract the ontologically defined meaning from the text in terms of who does what to whom, and when and where via CAMEO,[5] PETRARCH,[6] or some other coding framework (Schrodt 2006). Lastly, the disambiguated location names are evaluated to determine whether they represent the event-occurring location. While this step requires the completion of the two preceding steps, the enormous, extant body of work that has built up over the past decades of machine-coding of events has made the first two steps more manageable.

Named entity recognition in the context of geolocation involves determining which words in the given sentences are location names. In principle, the task of capturing location names from texts can be done easily by using a dictionary. In practice, however, developing a dictionary that is sufficiently comprehensive for such a task may be challenging. To begin, the geographic boundary of the texts being analyzed may be unclear, given that the domain of many event data

[2]Upon publication a repository of our project will be available at:
texttthttps:github.com/(author ID hidden)/LostInSpace.

[3]For the OEDA project, which is still in early development, see: `http:phoenixdata.org`.

[4]While there are twenty action verbs in the CAMEO ontology, verbs such as "appeal," "consult," or "yield" do not yield as many events in the data as "protest," "assault," or "fight" do. Also, the data of such events are not in the settings of interstate or intrastate conflicts. Hence we selected "protest" and "fight" data as test cases. Our method, however, is applicable to all the other CAMEO event types.

[5]For more information on CAMEO ontology, please see: `http:eventdata.parusanalytics.com/data.dir/cameo.html`.

[6]See `https:github.com/openeventdata/petrarch`.

is the entire world. Further, because conflict events often spread to new and rural places, texts may include location names not defined in the gazetteer. Still further, a location name may be written in multiple forms, requiring the dictionary to comprise every variant for each location. This commonly occurs when a foreign location name is transliterated into another language such as English. For instance, the transliterations *dei ez-zor*, *Deir-al-Zour*, *Dayr al-Zawr*, and *dei ez-Zour* all refer to the same province in Syria. Without specific dictionaries and correspondences, this is difficult to determine automatically.

Further complicating matters, news articles often use nearby landmarks to indicate the location, in lieu of using the official names. The 2014 Ukrainian revolution, for example, was often described as having taken place at Mariinsky Park, rather than in Kiev. The same is true for the so-called Martyr Square (Tahrir Square) and its role as a site of protests during the Egyptian revolution of 2011. We encountered this problem in our data set as well. For example, "U.N. attack helicopters whirred overhead as armoured personnel carriers ploughed through forests in Virunga National Park [a park in North Kivu province] in Democratic Republic of Congo . . . ".[7]

The dictionary approach can be complemented by the part-of-speech (POS) tagging method that grammatically parses sentences in the text and classifies each word into various categories such as persons, organizations, and location names. The technique can identify landmarks (for example, Mriinsky Park, Martir Square, or Virunda National Park) as locations even if they were not pre-defined in the dictionary. Currently, there are a number of open source systems available for named entity recognition. Typical software programs for this task include the Stanford Named Entity Recognizer (part of Stanford NLP), Apache Open NLP algorithm, and MIT Information Extraction (*MITIE*), developed by MIT's Lincoln Laboratory. The parsing stream for the OEDA pipeline, for example, combines POS tagging and the dictionary approach.

As shown in Figure 1, in OEDA the parser calls the Stanford CoreNLP, which returns to the Mongo database parsed sentences with parts of speech tagged. These parsed stories then are coded via the PETRARCH ontology. Only then is geolocation undertaken, using calls to the CLIFF geolocation software. While this method extracts an extensive list of location words, coding all country-relevant location words as the correct event location introduces a different problem.

Entity disambiguation is a nascent field of research aiming at determining the true location of the referent location word. (Han *et al.* 2011; Rao *et al.* 2013; Bunescu & Pasca 2006; Cucerzan 2007). For instance, "Durham" in the sentence,"the group moved to the intersection of Duke and Chapel Hill streets near the Durham Police Department headquarters" (Jul. 21st, 2016. CBS News Carolina), would most likely to refer to a city in North Carolina, U.S., while the same word in the sentence "Paul Collingwood commits for another year to Durham" (Jul. 26th, 2016. AFP) would most likely to refer to a city in England.

Different approaches exist for assigning location name strings to the referent location words. The co-occurences of location names, i.e. which location words frequently appear together in the corpus, could be modeled and used to link the name strings and the true locations (Han *et al.* 2011). Similarly, the *Mordecai* algorithm links the extracted location names to the locations defined in a gazetteer, by adopting *Word2vec* model (Mikolov *et al.* 2013) that calculates the co-occurrences of the words and quantifies the contexts in which specific words appear.[8] A well-trained corpus archive should be able to show words such as "Duke" and "Chapel Hill"

[7] ICEWS story ID: 4590482, DRC data

[8] Mordecai is described at `https:github.com/openeventdata/mordecai`.

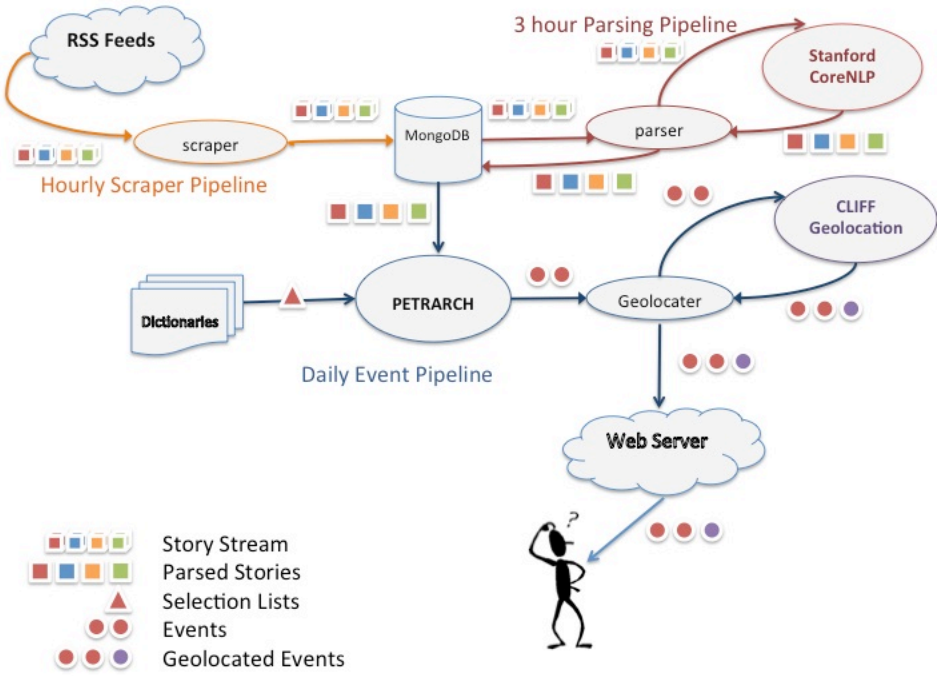

Figure 1. Schematic Representation of OEDA Pipeline, circa June 2016

appear commonly with "Durham" when "Durham" is the city name in North Carolina. While the disambiguation task is not simple, well-defined dictionaries may suffice these techniques when processing news articles that have clearly defined country bounds. But for projects that contain new and unknown sub-national location names, building an extensive dictionary is a daunting task, especially if the locales are not named in English.

Finally, geolocating events (identifying the location of the event described in a document) is an objective for many scholars, particularly those who intend to collect and build original databases from text corpora, be they news articles, congressional records, campaign speeches, party constitutions, or twitter feeds. While automating this task will aid many, the research avenue in this topic is still under development. One of the most commonly used methods is building location dictionaries and capturing location names. For instance, the principal investigators of the Project Civil Strife (PCS) data used three location dictionaries—"cities," "regions/provinces," and "others"—and coded all captured location names as the place of event (Shellman 2008). This approach, however, also includes irrelevant places as event locations. In our China data, for example, a total of 614 location words were captured from 250 news articles but only half of them (314 correct location words) are actual event locations. Given that a substantial number of location words are incorrect event locations, the automated event data community needs a better coding scheme that can reduce the error rates. Many have noted that this problem is yet to be solved. We discuss the remaining challenges in detail in the ensuing section.

3. Challenges in Geolocating Events

Selecting the correct location word among all captured locations is a difficult problem. In the data we examined, nine out of ten news articles contain multiple locations. Some of these are specious locations, indicating for example the location of news agencies, the location of a similar, often previous event, or the current location of a reporter. The occurrence of multiple location names not only introduces noise into the data, but also escalates the difficulty in automating the task of geolocation. Noting this difficulty, Schrodt & Yonamine (2012) states that "the main challenge is to empirically determine which place name should be assigned to the specific event, especially when multiple events and location names occur in a single article" (page 19).

Multiple approaches have been devised. One approach, adopted by ICEWS, is to code the location name that is the nearest to the action verb (identified by the TABARI coder[9]) in the text. Under such a scheme, a location word that is distant from the action verb is automatically discarded. Although the rationale for the algorithm sounds intuitive, errors frequently occur because action verbs are not always adjacent to the names of the event location.

The current OEDA data deployed uses a java-based web service named CLIFF.[10] This approach selects the most likely place as the "focus" location of the article, based on the frequencies of mentions and the order of appearance (D'Ignazio *et al.* 2014).

Notably, both ICEWS and OEDA assume that one correct event location exists per article. Yet, that assumption does not always hold. Over 30% of Syrian stories we examined contain multiple true locations, and in China and the DRC data, these numbers are about one-third to one-half this amount. For the other 175 countries in the world, these ratios, to our knowledge, are not yet known, but we can assume that the ratio is greater than zero. As the article in Table 4 demonstrates, single stories frequently contain multiple true location names. By assuming that a single location word exists per article as "the most appropriate location" (Lautenschlager *et al.* 2016), such a coding rule misses many true event locations, thereby hindering the accuracy of the coded location names. Moreover, a researcher who analyzes an event data under these approaches may misinterpret the data, for example, concluding that protests in China are concentrated in the capital, Beijing. However, even when there is a single location, the OEDA and ICEWS geolocation still will make many mistakes.

On the other hand, coding all location names in a text as the correct locations, as the PCS project has done, reduces false negatives but increases false positives. Hence, the optimal approach would determine which set of location words in the article is more likely to be the correct ones, in addition to relaxing the assumption that only a single location word represents each news article.

4. Categorizing Multiple Locations

In examining articles with multiple locations, we observed four mutually exclusive types of location words: 1) event-relevant and event-occurring, 2) event-irrelevant and event-occurring, 3) event-relevant and non-event-occurring, and 4) event-irrelevant and non-event-occurring. We define event-relevant locations as those locations that are part of the main description of the

[9]For the TABARI coder see: `http:eventdata.parusanalytics.com/software.dir/tabari.html`.

[10]CLIFF documentation: `https:github.com/openeventdata/phoenix_pipeline` and `http:cliff.mediameter.org`

Table 1. Location Word Type 2 (Incorrect) Vs. Type 1 (Correct); ICEWS story ID: 17929606, DRC data; and OEDA story ID: 2054559 v0.1.0, Syria data

	event-**ir**relevant and event-occurring Vs. event-relevant and event-occurring
Eg. 1	The Syrian Observatory for Human Rights monitor said that 15 civilians, among them 11 children, were killed in the attack on the Bab al-Nayrab neighborhood in the south of Aleppo.. . . Meanwhile, a ceasefire has been agreed in the town of Daraya, allowing 700 rebel gunmen safe passage to the northern province of Idlib and allowing 4,000 women and children to escape to shelters outside the town.
Eg. 2	1.. . . The government denounces a named refugee camp near Goma that was attacked by M23 soldiers. . . 6. Prime Minister Ponyo addressing an opening session of a seminar on agricultural sector in Kinshasa today.. . .

event of interest, i.e. all locations that are key to the narrative of the event of interest. Event-occuring locations refer to all locations where events occurred regardless of whether the event is the event of interest. Thus, the first category, *event-relevant and event-occurring*, refers to the locations where events occurred while the occurred events are within the scope of interest. We aim to detect this type of location words as the correct ones.

Regarding the second category, a small portion of articles contain *event-irrelevant and event-occurring* location words in our data. Such could occur when the raw texts contain news summaries of events that are not of interest. Table 1 shows examples of articles that contain both event-irrelevant and event-occurring (Idlib and Kinshasa) as well as the event-relevant and event-occurring (Aleppo and Goma) location words. The first example is from an article that describes a rebel attack event involving 15 civilian causalities and then describes a ceasefire agreement, an event not of interest. Sometimes, news articles contain summaries of completely unrelated events as in the second article, which consists of six reports that summarize events that occurred in Syria on that day.

The third type is *event-irrelevant and non-event-occurring* locations. Some of the most commonly observed event-irrelevant and non-event-occurring location names refer to the location of the news agencies and spokespersons. For example, Beijing in the first article in Table 2 indicates where the story was being written, but coded as the actual event location in ICEWS. We suspect that this is because the location word that refers to the reporting location is the closest location name from the action verb (strike), and hence was mistakenly coded as the event location. The true event location in this article is Guangdong.

Reporting locations often appear in the first line of the article. Discarding the reporting locations can alleviate the problem to a certain extent, but the problem persists because reporting locations are frequently embedded in the middle of texts, as demonstrated in the second and the third articles in Table 2. Also, the event-irrelevant and non-event-occurring location words are embedded for other reasons, such as referring to the birthplace of someone being interviewed, as in: " 'we can mix in any society,' said Amar Aldoura from Damascus."[11]

[11]OEDA story ID: 1424875 v0.2.0

The *event-relevant and non-event-occurring* locations complicate matters even more. Journalists often provide the background of the event being described. They may recite locations of the stronghold of a rebel group, the province name to which victims fled, or the place where the perpetrators of incidents are being trialed. The articles in Table 3 are examples of stories containing both event-relevant but non-event-occurring (Orientale and Damascus) and event-relevant and event-occurring locations (North Kivu and Daraa). In the first article, Orientale is a place to which the rebel leader was heading, so the location word is mentioned as part of the description of the rebel attack. But the actual attack was in North Kivu. The writer of the second article mentioned Damascus to describe a goal that the rebel group wishes to achieve. The actual attack was in Daraa.

To tackle the issue of creating the exhaustive list of location words that should be considered, we combine existing named-entity recognition, POS tagging, and entity resolution (matching location strings referenced in a gazetteer) techniques. Furthermore, based on the assessment of the types of multiple locations, we have come to the conclusion that each location word should be evaluated and determined whether it is an event-relevant and event-occuring location. For determining the boolean status (true event location or not) of each captured location word in the exhaustive list, we adopt a classification approach, which we discuss more in the next section. A sophisticated algorithm would distinguish the correct locations from the incorrect ones by filtering out the *event-irrelevant* locations, as well as *non-event-occurring* ones.

Table 2. Location Word Type 3 (Incorrect) Vs. Type 1 (Correct); ICEWS story ID: 18141520, China data; ICEWS story ID: 10672170, DRC data.

event-**ir**relevant and **non**-event-occurring Vs. event-relevant and event-occurring	
Eg. 1	Beijing, Nov 19, 2011 (AFP) - More than 7,000 workers went on strike at a southern Chinese factory... Dozens of workers were injured on Thursday as police tried to break the strikers' blockade of the main road in the factory town near Dongguan in Guangdong province...
Eg. 2	The clashes, which started at Luozi and to Seke Banza [towns in Bas-Congo Province]... Speaking to reporters in Kinshasa, the parliamentarian said that the clashes had left at least 100 people dead and many more nursing serious injuries since last Friday.
Eg. 3	The protesters gathered outside the office of Southern Weekly in Guangzhou, capital of southern Guangdong province, on Monday calling for media freedom, a taboo subject in the country, holding banners and chanting slogans.... A foreign ministry spokesperson in Beijing is reported to have said: "There is no so-called news censorship in China."

Table 3. Location Word Type 4 (Incorrect) Vs. Type 1 (Correct); OEDA story ID: 1160025 v0.2.0, Syria data; ICEWS story ID: 18922379, DRC data.

	event-relevant and **non**-event-occurring Vs. event-relevant and event-occurring
Eg. 1	The mutineer general, Bosco Ntaganda and his men who have been on the run for several days now, exchanged "heavy" gunfire with the army in the night of 7 May and is heading toward the Virunga National Park [a park in Orientale province]. . . . "After four hours of exchange of heavy gunfire at Kibumba," a locality at the border of the Nyiaragongo and Ruthshuru territories, in the volatile province of North Kivu, "we were supported by shots from heavy weapons," stated the captain.
Eg. 2	Elsewhere in Syria, 51 rebel factions operating in the southern province of Daraa announced a campaign to wrest control of areas of Daraa city [capital of Daraa governorate] from the government. . . . SANA reported that an attack by terrorists had been thwarted, with fighter jets pounding rebel targets in surrounding villages. If successful, it would grant the rebels a rear supply base to mount operations on Damascus. . .

5. Classification Algorithms

For classifying location words either as correct or incorrect, various machine learning techniques could be used, such as the following: Neural Networks (Müller & Reinhardt 2012; Mehrotra *et al.* 1997; Cheng & Titterington 1994),[12] SVM (Cristianini & Shawe-Taylor 2000; Vapnik 1995; Cortes & Vapnik 1995),[13] random forests (Liaw & Wiener 2002; Breiman 2001),[14] AdaBoost (Freund & Schapire 1997), K-nearest neighbors (K-NN) (Dasarthy 1990), and naive Bayes (Zhang 2004; Murphy 2006). Of these, we employ three classifiers—artificial neural networks, SVM, random forests—and compare the performance of each.

The artificial neural network models the relationship between a set of input signals, the desired feature from the texts, and an output signal—whether a location word refers to the event location—using concepts borrowed from our understanding of how a human brain processes information from sensory dendrites through neurons while allowing the impulse to be weighted according to its relative importance (*model parameters*). Although the algorithm is notoriously slow, the artificial neural networks have more flexibility in terms of structures and parameters compared to other classifiers (Lantz 2013; Zurada 1992). But a potential downside of neural network model is that its prediction performances usually relies on a considerable amount of training data.

SVM refers to support vector machines. These were initially introduced for solving two-group classification problems where the data are mapped into a higher dimensional input space and construct an optimal separating hyperplane (Vapnik 1998; 1995). This approach is often viewed as superior to other machine learning algorithms, including neural networks, because the quadratic programming guarantees reaching the global optimum, which often leads to the larger

[12] See Zhang & Zhou (2006); Ng *et al.* (1997) for examples of neural networks applications in text analysis.

[13] Examples of SVM applications in text analysis can be found in Minhas *et al.* (2015); Tong & Koller (2001); T. Joachims (1998).

[14] See Fette *et al.* (2007) for examples.

overall classification accuracy (Li 2003; Vapnik 1998; T. Joachims 1998; Maroco *et al.* 2011). Furthermore, SVM models are typically less prone to over-fitting (Mukherjee *et al.* 1997). SVMs also provide a computationally efficient way to achieve a reasonably accurate model (Amami *et al.* 2012; Li 2003).

Finally, the random forests (or decision tree forest) model, championed by Leo Breiman (2001) and Adele Cutler (Breiman & Cutler 2007), combines the principle of bagging with random feature selection to add complexity to the decision tree models. After the ensemble of classification regression trees (hence the name forest) is generated, the model combines these trees' predictions. Because the ensemble uses only a small, random portion of the full feature set, the model can handle large data sets wherein the high dimensionality may cause other models to fail. Despite the difficulty in interpreting the results, it is an all-purpose approach that performs well on most problems (Lantz 2013).

These classifiers boast two primary strengths. The first is that they do not require prespecifying the type of relationship between the covariates and the response variable. They are powerful information extraction tools that can capture underlying relationships not explained by known structures (Jones & Linder 2015; Lantz 2013; Günther & Fritsch 1998; Beck *et al.* 2000). Second, these models achieve accurate prediction rates given large enough input sizes (Maroco *et al.* 2011; Hsieh *et al.* 2011; Beck *et al.* 2000). Lantz (2013) suggest that these algorithms are the most accurate state-of- the-art approaches, and make few assumptions about the data.

Some scholars oppose the use of the machine learning in fields that require substantive interpretations of the parameters (de Marchi *et al.* 2004) because they are "difficult to interpret" (Lantz 2013). Despite such pitfalls, the prediction performance of these models make them attractive for geolocation. Whether they produce interpretable results can not be determined *a priori*.

6. Building Dictionaries

Before classification can begin, correctly formatted text data with desirable features is required. We developed four types of dictionaries in order to preprocess the text data. First and foremost, a location dictionary for each country was compiled. The initial location lists were imported from Geonames, Wikipedia, and Google map. These dictionaries contained province names (standardized province and governorate names) and sub-province names (city, village and town names, spelling variations of both province and sub-province names, and frequently used famous location names) as two separate columns. To ensure that our location dictionary was as comprehensive as possible, we used an iterative process to build it. After the initial location dictionary was built, we went back to the text data and parsed sentences using MITIE. The parts of speech elements classified as location words were then sent to the Genomes API and the returned entity pairs that were not already in, but should have been in the dictionary, were added.

We also developed dictionaries for actors. While we imported the actor lists from ICEWS and OEDA data and manually modified them depending on the salient actors in each country, the entire process can be done manually. Without the prior knowledge about the events of the data at hand, one may resort to sentence parsers and build dictionaries iteratively as we did for the location dictionary.

For the relevant words dictionaries, we first imported action verb lists from the CAMEO ontology, on top of which our data sets were built. The verbs for the protest data included words such as "rally," "demonstrate," and "march." For the fight data, the verb list included "air-strike," "bomb," and "shoot." For both dictionaries, we then added key nouns that capture the context of the location sentence such as "bloodshed" and "casualty." Likewise, a dictionary including irrelevant words, such as "report" and "interview", was compiled. As in the process of building the other dictionaries, the relevant words dictionary does not have to depend on any existing ontology but we chose to adopt the pre-existing framework of CAMEO because those action verbs were used to collect the news articles in our data in the first place.

Finally, dictionaries containing generic words that are not data specific were compiled. The lists included names of news agencies (for example, AFP, AP, and CNN), directional words (southern, southeastern, . . .), the names of months (january, jan, february, feb, . . .), and days (monday, mon, . . .). All of these dictionaries were used as part of preprocessing.

7. Preprocessing the Texts

The literature on text analysis describes a few common preprocessing steps. Following D'Orazio *et al.* (2014), we first removed punctuation and special characters from the text data that contain sentences with location words. We next converted all sentences to lower letters to avoid confusion in recognizing word patterns. We then removed stop words in English(Shellman 2008; Monroe *et al.* 2008). The stop words list was imported from the Stanford NLP Group, but we modified it to exclude prepositions related to locations, such as "in", "at", and "from". Next, we performed stemming (Grimmer & Stewart 2013), using Porter Stemmer (Porter 1980).[15]

In addition to the tasks performed prior to most text analysis projects, we also performed two additional text treatment tasks that are critical in our algorithm: 1) *homogenization* of location words and 2) *generalization* of texts using the dictionaries described above. As with many other text analysis projects, the accuracy of our algorithm depends highly on the quality of the pretreatment process.

The homogenization step is important because the use of location names in news articles is not always consistent with respect to the spelling of location names, particularly of those in non-English speaking countries. In addition to the transliteration issue, the different conventions of stating locations also complicate the process. For example, news articles by local agencies cite only city names while those by national or international agencies often indicate only province names. Accordingly, we used the location dictionaries to standardize these variations across data. For the sentences that contain only the lower level location words, we converted the administrative division names (city) to higher level ones (province/governorate) with the prefix of "sub-". For instance, "Fataki, Orientale" would be converted to "sub-orientale, orientale." In building the dictionaries, we used the administrative division at the time of the news reports. For instance, city A in year 2005, the year of the event, may be in province B, but in province C in year 2016. In such a case, we used province B as the correct province.

[15]We were careful to preserve important information. For instance, Porter Stemmer removes 'ing' at the end of each word, so we converted some province names such as "liaon" back to "liaoning".

As the last step in the preprocessing stage, we generalized the news texts using the aforementioned dictionaries of actors, relevant and irrelevant words (action verbs, key nouns, irrelevant verbs and nouns), numbers, dates and news agencies.[16] The purpose of this step is to ensure that the algorithm would recognize the following two N-grams as identical[17]: "33 people in Beijing" and "2000 people in New York".[18] More generalized sentence patterns are desirable because the approach aims to match patterns of phrases and sentences from different news articles. An example of a preprocessed text looks like the right side of Table 4.

Table 4. Example of Raw and preprocessed Text;ICEWS story ID: 25149588, China data

Raw text	Treated text
BEIJING, Dec 4 (AFP) – More than 500 people protested outside government offices after a man died under suspicious circumstances while in police custody, a human rights group and relatives said Thursday. It is one of two such cases – in northern Shandong province and southeastern Fujian province... earlier this year the beating death of a man in a police detention hall in southern Guangdong sparked widespread criticism....	NUMERAL ACTOR ACTION-VERB outsid ACTOR man ACTION-VERB suspici circumst ACTOR custodi ACTOR relat said DAY. NUMERAL case DIRECTIONAL shandong ADMIN DIRECTIONAL fujian ADMIN... earlier DATE ACTION-VERB ACTION-VERB man ACTOR detent hall DIRECTIONAL LOCATION spark widespread criticism....

8. Implementation of Automated Classifiers

To mimic the way human coders would parse sentences and retrieve the relevant information, we trained the machine to learn the collocation patterns of the correct and incorrect location words and then to predict the correctness of a new set of location words based on the collocation patterns of those new words. This approach of storing patterns and solving problems known as case-based reasoning is a common paradigm in automated reasoning and machine learning in which a reasoner solves a new problem by using a similar problem that has already been solved (Kolodner 1992; de Mántaras & Plaza 1997).

The implementation of our algorithm involves two stages: feature selection and model estimation. These stages require pre-treatment of the text data as illustrated in Figure 2. To describe the feature selection stage in detail, we take examples from the China data, which consist of 250 news articles on protest towards the government from 2001 to 2014.[19] On average, each

[16]The entries in these dictionaries were stemmed.

[17]Location names are generalized in the algorithm after a specific location name is collected. Hence, they should not be generalized during the preprocessing stage.

[18]If these two phrases are generalized in terms of numerals and location names, they become "numeral people in location".

[19]We removed duplicate reports.

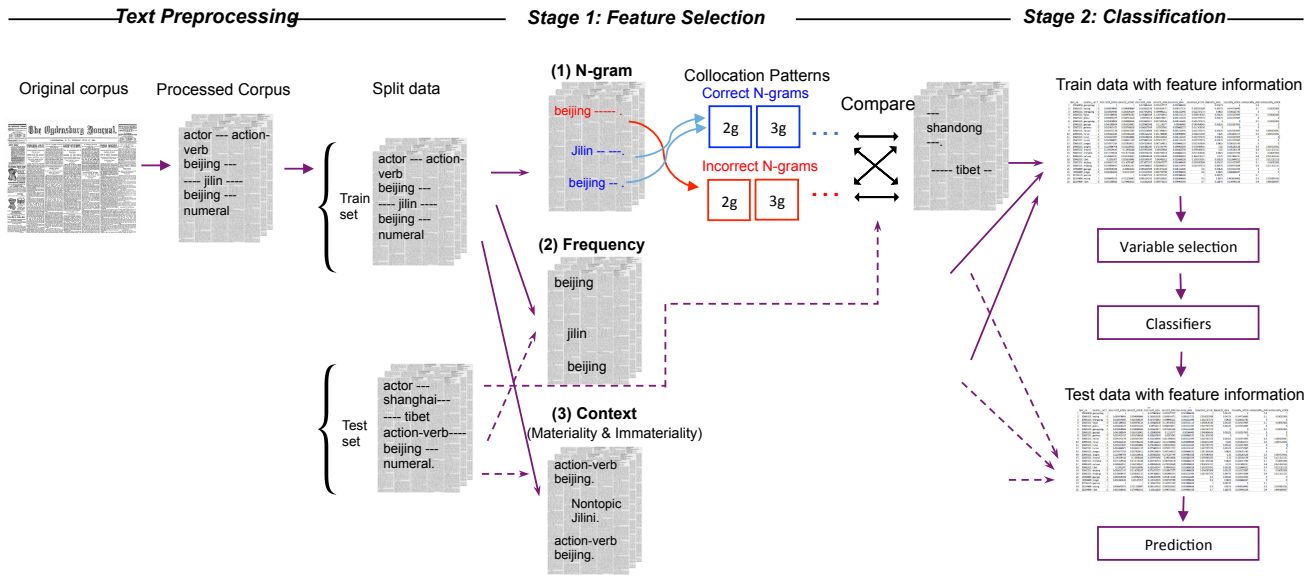

Figure 2. Flow Chart

article contains about 398 words before the preprocessing treatment and 284 after the treatment. For each validation score (of the nine results that are averaged and presented in Table 7), we randomly divided the treated articles into training and test sets.

What information then do we feed the machine to develop classifiers? Our goal is to differentiate event-relevant words from irrelevant ones and event-occurring words from non-event-occurring ones. Therefore, we select variables that can provide information about "event-relevance" and "event-occurrence." Specifically, 1) N-gram collocation patterns, 2) frequency of location words, and 3) context of the sentences that contain the location word are extracted from the news articles.

An N-gram is a sequence of N words. Collections of N-grams are known to provide valuable information about each word in a phrase, taking into account the complexity and long distance dependencies of languages.[20] In a sentence "Factory workers protested", an N-gram of order 2 (or bigram) is a two-word-sequence of words (for example "factory workers", and "workers protested") while an N-gram of order 3 (or trigram) is a three-word-sequence of words (such as "factory workers protested"). Given that the collocation patterns in which the event-occurring location words appear differ from those of the non-event-occurring collocations, the N-gram patterns are able to provide the contextual information of event-occurence to our classifiers. We thus compare the frequencies of each N-gram in correct and incorrect corpus and compute the relative frequencies, respectively.

Some examples of N-grams collected from one of the training sets are shown in Table 5. From the raw text on the left in 4, "LOCATION MONTH"—the bigram for Beijing—will be stored in the

[20]For more information on N-gram, see Ch. 4 in Jurafsky & Martin (2009).

incorrect bi-gram corpus while “DIRECTIONAL LOCATION”, “LOCATION ADMIN”, —the bigrams of Shandong—and “DIRECTIONAL LOCATION’, “LOCATION ADMIN”—the bigrams of Fujian—will be stored in the correct bigram corpus. Some examples in the bi-grams (N-grams of N=2) of correct location words from the training set in the China data include “LOCATION ADMIN" (frequency: 17), “LOCATION ACTION-VERB" (frequency: 15), and “outsid LOCATION" (frequency: 15). These bi-grams were extracted from sentences such as the following: “The *Guizhou provincial* government deployed thousands of police",[21] “Workers at IBM Systems Technology Company (ISTC) in *Shenzhen are protesting* since March",[22] and “500 villagers had been protesting *outside the Qingdao* naval base".[23] Two of the most frequent bi-grams in the incorrect corpus are “LOCATION MONTH", and “LOCATION ACTOR". They are from phrases such as “*BEIJING, Dec* 3, 2007 (AFP)",[24] “*Shandong farmers* protested. . . ".[25] Table 5 shows the top 10 most frequent bi-grams for both correct and incorrect location words in one of the training sets.

Table 5. Examples of Collocation Patterns (n=2)

	Correct	Freq	**Incorrect**	Freq
1	of SUB-LOCATION	69	LOCATION MONTH	27
2	of LOCATION	33	LOCATION ACTOR	20
3	in LOCATION	30	to LOCATION	19
4	DIRECTIONAL LOCATION	24	LOCATION SOURCE	18
5	in SUB-LOCATION	24	from LOCATION	16
6	LOCATION ADMIN	17	by SUB-LOCATION	15
7	LOCATION and	17	link LOCATION	14
8	SUB-LOCATION near	16	NONACTION-VERB SUB-LOCATION	12
9	LOCATION ACTION-VERB	15	to LOCATION	11
10	outsid LOCATION	15	near LOCATION	10

Then we compare the captured collocation patterns, consisted of location words and their neighboring words, to the correct and incorrect N-gram lists. In the texts in Table 6, for instance, “heilongjiang” and “beijing” would be captured. While creating covariates for “heilongjiang”, the N-gram collocation patterns, such as “of heilingjiang”, “at sub-heilongjiang”, and “in sub-heilongjiang”, are converted to “of LOCATION”, “at sub-LOCATION”, and “in sub-LOCATION”. For “beijing”, the N-gram collocation patterns such as “to beijing”, “beijing therefor”, and “in beijing” would be converted to “to LOCATION”, “LOCATION therefor”, and “in LOCATION”. These generalized N-gram patterns are compared to the correct and incorrect pattern lists (compiled from the training set) that looks like the list in Table 5. Then the N-gram pattern feature is converted to numeric values in two ways. The first N-gram variables compute the ratio the collocation patterns comparing both the stored correct and incorrect pattern lists. The other N-gram variables

[21]Story ID: 35682875, ICEWS China data
[22]Story ID: 32977476, ICEWS China data
[23]Story ID: 32852391, ICEWS China data
[24]Story ID: 22997344, ICEWS China data
[25]Story ID: 21984369, ICEWS China data

reflect how many of these collected patterns can be matched to the most frequent patterns in each list.[26]

The second type of variables, the frequencies of location words, provide the information about the relevance of a particular location word. Assuming that the news articles in the data are well-sorted and contain articles mostly pertinent to the research interest, the set of location words that are mentioned several times should have higher chances of being correct, compared to the ones with low frequencies (D'Ignazio *et al.* 2014).

Table 6. Illustrative Example, ICEWS story ID: 25149588, China data with Province name added by the author.

Raw text	Treated text
About 500 angry textile workers blocked a railway line in northeastern China on Monday demanding unpaid wages and unemployment pay from the government, said railway employees who saw the protest. There were four to five hundred of them blocking the railway. They stood on the railway, but were later dispersed," said a man who worked at a railway station in Jiamusi [town in Heilongjiang] the northeastern province of Heilongjiang. "The train to Beijing was therefore delayed five to six minutes in leaving," he said. The protest was similar to one in December when more than 2,000 workers from the same bankrupt textile plant blocked a rail line and cut traffic on an airport highway, accusing company officials of embezzling their social security payments. An employee at Jiamusi [town in Heilongjiang]'s main train station said city and railway police went to persuade them to leave and protesters dispersed after about 20 minutes, he said. The plight of disgruntled workers laid off from bloated state-owned firms, like those in Jiamusi [town in Heilongjiang], is getting top billing at the annual two-week session of the National People's Congress, or parliament, meeting in Beijing due to end on March 18.	N angri textil ACTOR ACTION-VERB railway line in DIRECTIONAL ACTOR on DAY demand unpaid wage and unemplo ACTOR pay from gover ACTOR NONTOPIC railway ACTOR saw ACTION-VERB. N to N of ACTION-VERB railway. stood on railway later dispers NONTOPIC man work at railway station in sub-heilongjiang in DIRECTIONAL ADMIN of heilongjiang. train to beijing therefor delay N to N minut in leav NONTOPIC. AVERB similar to N in MONZ N ACTAR from ACTAR upt textil plant ACTION-VERB rail line and cut traffic on airport highway NONTOPIC ACTOR of embezzl social secur pa ACTOR s. ACTOR at sub-heilongjiang main train station NONTOPIC ADMINN and railway ACTOR went to persuad to leav and ACTOR dispers N minut NONTOPIC. plight of disgruntl ACTOR laid from bloat state own firm like in sub-heilongjiang get top bill at annual N week session of nation peopl ACTOR par ACTOR meet in beijing due to end on MONTH N...
ICEWS location: Beijing—Correct location: Heilongjiang	

[26]We used the top 50% of N-gram patterns in terms of frequencies. For the size of our data, the most frequent correct and inccorrect N-gram lists included 10 to 20 collocation patterns.

Testing the context, sometimes called materiality, of the sentence that contains location words is another way of capturing relevant location words. The idea is that, if the sentence contains more action verbs and key nouns, the location word in that sentence is highly likely to be relevant. Likewise, a location word in a sentence with "report" or news agency names may be less likely to be relevant.

Finally, we designed the data so that it can account for the variations at the article level as well as the data level, assuming that 1) some location words are more correct than others in each article and 2) some articles contain location words that are collectively more likely to be correct or incorrect altogether. In other words, these variables are calculated in relative terms within the article and data levels. This means that for the within article level variables, the location word with the largest value in that article has the value of 1. At the data level, only one location word with the largest value within the data has the value of 1. For example, the relative within article ratio of frequency for Heilongjiang in the example in Table 6 would be 1 while that for Beijing would be 0.67. Table 6 also shows the first and the second sentences containing "heilongjiang" include a irrelevant word "said", converted as "NONTOPIC", and therefore, the within-article materiality ratio for Heilongjiang and Beijing would be 0 and 0 while the immateriality ratio for the two would be 1 and 0 respectively. All the positive values would be much smaller in the within data ratios.

To extract the above mentioned features, we start with the training set, which consists of two thirds of our text data. We first capture all location words that match the list in our location dictionary, and determine whether the location word falls into the correct category or the incorrect category, based on human coding. Once the recognized location word is determined as either correct or incorrect, they are stored separately for correct and incorrect corpora.

Once the corpora of correct and incorrect N-grams are created, we compute the N-gram pattern information (N being the range specified[27]), the frequencies of mention, and the materiality of the location sentences, in terms of both within article and data level ratios. The final data generated would contain the dependent variable indicating whether the particular location word is correct (Y=1) or not (Y=0), and the covariates of the frequencies of each N-gram and the three covariates mentioned above.

Figure 3 shows the first thirteen rows and parts of covariates of the data generated using the Chinese news articles. The number of rows of the data equals the number of total province names appearing in all news articles. The first column represents the unique story IDs from ICEWS and the next column contains all of the location words in the article. The Y variable shows whether the location word is correct or not, based on the human coders' judgment. In the data shown, the first row represents the story 1517019 from the ICEWS data, the example in Table 6. The article contains two location words, 'beijing' and 'heiliongjian', of which the second is the correct event location. The next four location words in rows three through six are from a single article, ICEWS story 16963437. Of these, only 'sichuan' is the correct event-relevant and event-occurring location.

The covariates with the suffix "article" are the relative ratios within articles. Location words within articles that do not contain any other locations are therefore assigned the value of one. The covariates with the suffix "data" represent the relative ratios within the data. Correct and

[27]We computed this frequency rate for each location word for N-grams of two to seven.

	text_id	location_name	Y	incorrect2_article	correct2_article	incorrect2_data	correct2_data	incor.matched2_a	cor.matched2_ar	incor.matched2_da	cor.matched2_da	occurence_article	occurence_data	materiality_article	materiality_data	imateriality_article	imateriality_data
1	1517019	beijing	0	0.627906977	0.234899329	0.127962085	0.080831409	1	0.166666667	0.045454545	0.076923077	0.333333333	0.1	0	0	0.142857143	0.1
2	1517019	heilongjiang	1	1	1	0.203791469	0.344110855	0	1	0	0.461538462	1	0.3	1	0.058823529	0.857142857	0.6
3	16963447	jiangsu	0	0.292682927	0.258883249	0.056872038	0.058891455	0	0.25	0	0.076923077	0.333333333	0.05	0.333333333	0.058823529	0	0
4	16963447	liaoning	0	0.317073171	0.263959391	0.061611374	0.060046189	0	0.25	0	0.076923077	0.333333333	0.05	0.333333333	0.058823529	0	0
5	16963447	sichuan	1	1	1	0.194312796	0.227482679	0	1	0	0.307692308	1	0.15	0	0	0	0
6	16963447	zhejiang	0	0.341463415	0.269035533	0.066350711	0.061200924	0	0.25	0	0.076923077	0.333333333	0.05	0.333333333	0.058823529	0	0
7	22834745	guangdong	1	1	1	0.047393365	0.106235566	0	1	0	0.076923077	1	0.05	1	0.235294118	1	0.1
8	25942100	beijing	0	0.384615385	0.078651685	0.023696682	0.008083141	1	0	0.045454545	0	1	0.05	0	0	0	0
9	25942100	liaoning	1	1	1	0.061611374	0.102771363	0	1	0	0.153846154	1	0.05	1	0.117647059	1	0.1
10	22285172	beijing	1	0.178571429	0.088571429	0.047393365	0.035796767	0	0.2	0	0.076923077	0.25	0.05	0.2	0.176470588	0.5	0.2
11	22285172	guangdong	0	0.928571429	0.742857143	0.246445498	0.300230947	0	0.8	0	0.307692308	0.75	0.15	0	0	0	0
12	22285172	hunan	1	0.696428571	0.594285714	0.184834123	0.240184758	0	0.6	0	0.230769231	0.5	0.1	0.333333333	0.294117647	0	0
13	22285172	shandong	1	1	1	0.265402844	0.404157044	1	1	0.045454545	0.384615385	1	0.2	0.466666667	0.411764706	0.5	0.2

Figure 3. Data Generated (China)

incorrect N-grams of two to seven, frequencies, and materiality variables are constructed in this manner.

In Stage 2, with the data (of the training set) generated, we fit the artificial neural networks, the support vector machine, and the random forests models. Using the random forests Recursive Feature Elimination (RFE) algorithm, we selected the variables of which the combination yields the highest accuracy rates in the training data.[28] The parameters of each classifier were adjusted to get the optimal result. In random forests, the number of trees was set to 1000 and the kernel radial was set for SVM. The artificial neural networks model was tuned in each iteration, selecting automatically the best decay rate and the number of dendrites in the hidden layer.

The estimated parameters were then used to predict the boolean status of each location word in the test set. For the example in Table 6, the average predicted probabilities for Beijing as the correct location is 6% (neural net), 9% (SVM), and 27% (random forests) and those for Heilongjiang as the correct location is 97% (neural net), 75%(SVM), and 98% (random forests), making the predictions for *both* location words correct.

9. Results

Table 7 summarizes the performance statistics of various methods including our own classification approach. The columns represent each method and the rows indicate the data sets used. All numbers are rounded up. For the classification algorithms, we performed three 3-fold cross validations, thus the scores in the columns of neural net, SVM, and random forests are averages of nine iterations in total. The dictionary column represents the accuracy rate of the dictionary method that codes all captured location words as correct event locations.

Table 7. Validation Results

Datasets	Classification (NNet)	Classification (SVM)	Classification (R.Forests)	Dictionary
China (ICEWS, Protest)	73	75	73	51
D. R. Congo (ICEWS, Fight)	81	84	84	59
Syria (OEDA, Fight)	74	73	75	57

As Figure 4 shows, the accuracy rates across models do not vary much with about 3% maximum difference. While the performance of our algorithm is consistently high across classifiers, the highest accuracy rates in each data set were produced by SVM for the China and DRC data, and by random forests for the Syria data. However, the results vary across data sets, with the

[28]The N-gram patterns were the most powerful variables consistently.

highest rates in the DRC data. This difference comes from the number of true locations in the article. The accuracy rates for location words in the subset consisting of only articles with one correct location range from 86% in China to 90% in the DRC.

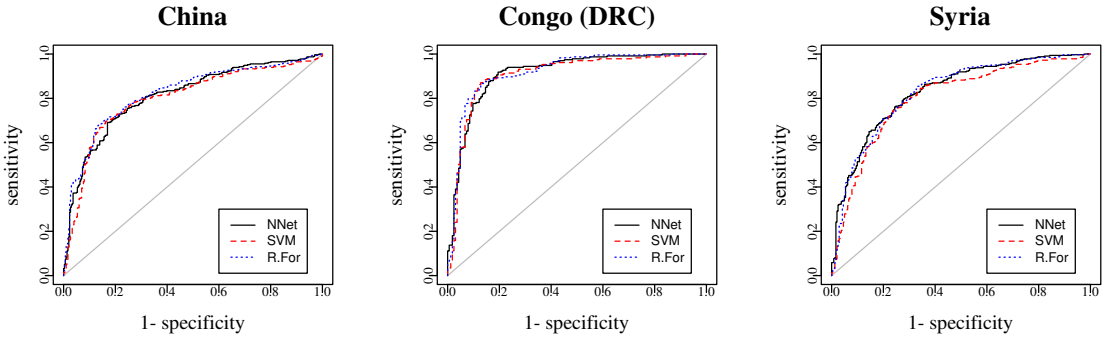

Figure 4. Receiver operating characteristic Curves

We have compared the true positives of our results to the current machine-coded data. These are the location words that each algorithm classifies as the actual event locations. Figures 5 to 7 compare the results to the ground truth which is plotted on the left and the current ICEWS or OEDA locations on the right. The sizes of bubbles represent the relative shares of events and the colors represent frequencies with the legends on the far right. In the China data, our algorithm misses eleven protests in Sichuan, but in all other provinces the differences are single digits. On the other hand, ICEWS codes Beijing as the event location more than 30 cases than the ground truth and misses more than 30 protest cases in Guangdong province alone.

Results in the DRC are accurate regardless of the choice of classifiers and the best performance is around 85%. This is true in part because the events in the DRC do not typically include a large number of locations. The civil conflicts which show up in the fight category in the DRC are concentrated in a small number of areas. By comparision, protests in China are not only in a much larger country, but are in a wide variety of locations. Accordingly, stories about China have many more location words per story, and are harder to correctly identify than is the case in the DRC. Figure 6 shows that the bubbles of the human-coded map on the left and the machine-coded map in the middle are nearly identical. Compared to the human coded locations, the locations coded by ICEWS model are correct at around 67% with over 30 under-reporting cases of fight in North Kivu and over 20 over-reporting cases in Kinshasa.

In the Syria case, our overall predicted event locations also look very similar to the ground truth while OEDA not only misses over one-half of the true event locations (129 NAs in 250 news articles), but also includes event locations that are not in Syria such as Beirut (three events), Illinois (one event), Moscow (one event), New Jersey (one event) and Pennsylvania (four events). Compared to the human coded locations, event locations in OEDA data are correct 31% of the time.

Overall, the performance of our classifiers is strong, improving the accuracy rate by as much as 25% from the dictionary approach. Furthermore, even if the accuracy rate is not 100%, because our algorithm evaluates each location word, it does not symmetrically miss or favor certain locations as the current ICEWS and OEDA algorithms do. Hence, the visualized results seem very close to the ground truth.

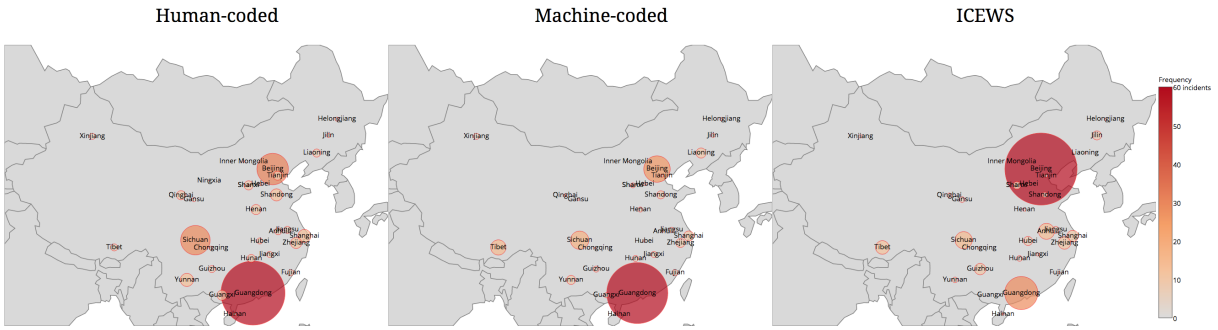

Figure 5. Protest Frequencies in China

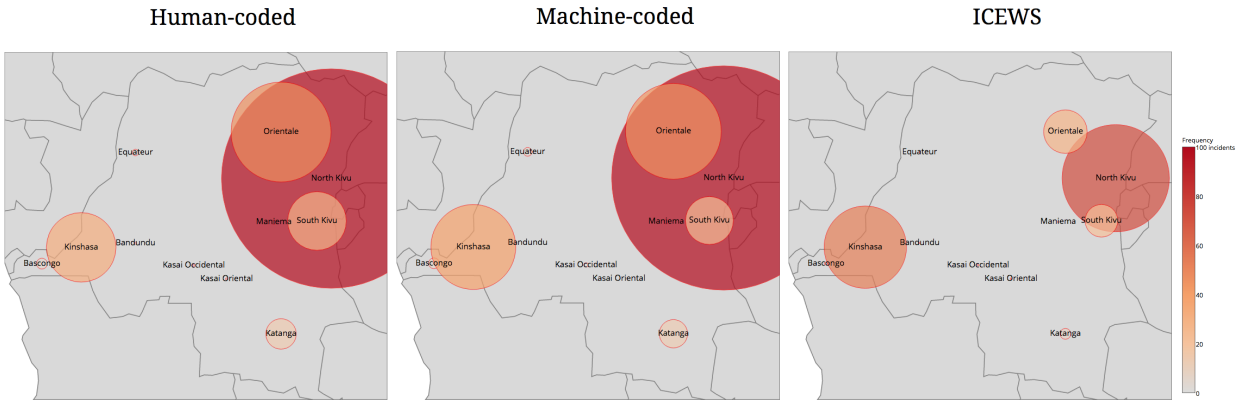

Figure 6. Fight Frequencies in Democratic Republic of Congo

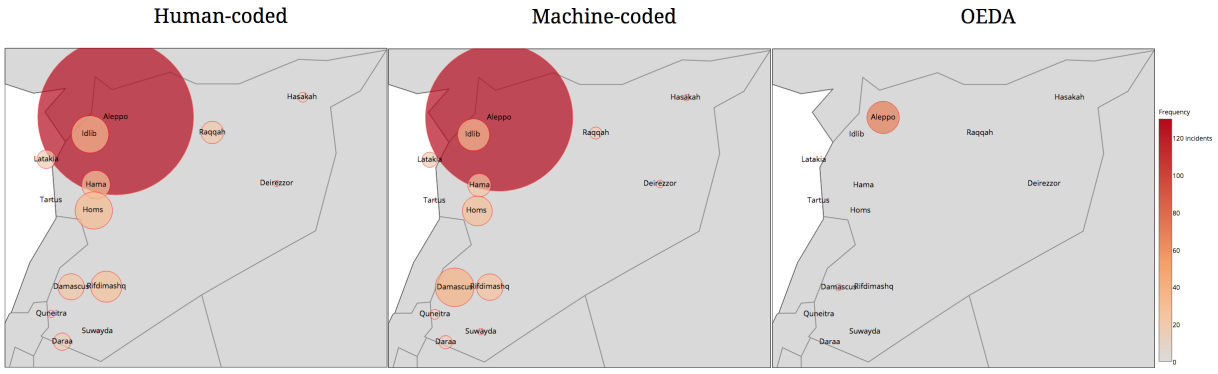

Figure 7. Fight Frequencies in Syria

10. Conclusion

We examined some problems associated with current geolocation methods employed in existing machine-coded event data. Locations of events contain valuable information that is of interest to many scholars and policy makers. To address discrepancies in geolocation between automated and human coders, we developed a supervised machine learning algorithm that filters out event-irrelevant locations as well as non-event-occurring ones. Departing from the assumption that one correct location exists in a news article, we evaluate each location word. By

doing so, we diverge from algorithms that are systematically biased towards certain locations such as the capital of a country and locations that appear frequently in the corpus. Using human coded ground-truth, we demonstrate that this approach is superior to extant approaches in the cases we have studied.

Interested scholars can extend the current work to a wider range of event ontologies and locations. While we have studied only a few countries, the protocol we developed may aid others who are interested in different countries to geolocate extant event data more accurately. Others who wish to extract location information from structured text data written in formal language, such as the United Nations reports on Children and Armed Conflict (`https:childrenandarmedconflict.un` or Amnesty International country reports (`https:www.amnesty.orgen/latest/research/2016/02/annual-`) can utilize our (open source) protocol—available upon publication—to create new event data streams in which the events are geolocated.

References

Amami, Rimah, Ayed, Dorra Ben, & Ellouze, Noureddine. 2012. An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel Recognition. *International Journal of Intelligent Information Processing*, **3**(1).

Arguello, Jaime, Callan, Jamie, & Shulman, Stuart. 2008. Recognizing Citations in Public Comments. *Journal of Information Technology and Politics*, **5**(1), 49–71.

Beck, Nathaniel, & Katz, Jonathan N. 1995. What to Do (and Not to Do) With Pooled Time-Series Cross-Section Data. *American Political Science Review*, **89**(3), 634–647.

Beck, Nathaniel, King, Gary, & Zeng, Langche. 2000. Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review*, **94**(1), 379–389.

Boschee, Elizabeth, Lautenschlager, Jennifer, O'Brien, Sean, Shellman, Steve, Starz, James, & Ward, Michael D. 2015 (March). *ICEWS Coded Event Data*. http://dx.doi.org/10.7910/DVN/28075 Harvard Dataverse Network [Distributor] V1 [Version].

Breiman, Leo. 2001. Random Forests. *Machine Learning*, **45**(1), 5–32.

Breiman, Leo, & Cutler, Adele. 2007. *Random Forests-classification Description*.

Bunescu, Razvan, & Pasca, Marius. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. *Pages 9–16 of: Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*.

Cardie, Claire, & Wilkerson, John. 2008. Text Annotation for Political Science Research. *Journal of Information Technology & Politics*, **5**(1), 1–6.

Cederman, Lars-Erik, & Gleditsch, Kristian Skrede. 2009. Introduction to Special Issue on "Disaggregating Civil War". *Journal of Conflict Resolution*, **53**(4), 487–495.

Cheng, Bing, & Titterington, D. Michael. 1994. Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, 2–30.

Cortes, Corinna, & Vapnik, Vladimir. 1995. Support-vector Networks. *Machine learning*, **20**(3), 273–297.

Cristianini, Nello, & Shawe-Taylor, John. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, NY: Cambridge University Press.

Cucerzan, Silviu. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Pages 708–716 of: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech

Republic: Association for Computational Linguistics.

Dasarthy, Belur V. 1990. *Nearest Neighbor Pattern Classification Techniques*. Hoboken, NJ: IEEE Computer Society Press.

de Mántaras, Ramon López, & Plaza, Enric. 1997. Case-Based Reasoning: An Overview. *AI Communications*, **10**(1), 21–29.

de Marchi, Scott, Gelpi, Christopher, & Grynaviski, Jeffrey D. 2004. Untangling Neural Nets. *American Political Science Review*, **98**(02), 371–378.

D'Ignazio, Catherine, Bhargava, Rahul, Zuckerman, Ethan, & Beck, Luisa. 2014. Cliff-Clavin: Determining geographic focus for news articles. *In: title = KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, New York, USA: Association for Computing Machinery, for Association for Computing Machinery.

D'Orazio, Vito, Landis, Steven T., Palmer, Glenn, & Schrodt, Philip. 2014. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines. *Political Analysis*, **22**(2), 224–242.

Easton, Kent, Kaiser, Kai, & Smoke, Paul. 2011. *The Political Economy of Decentralization: Implications for Aid Effectiveness*. Washington D.C.: The World Bank.

Fearon, James D., & Laitin, David D. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review*, **97**(1), 75–90.

Fette, Ian, Sadeh, Norman, & Tomasic, Anthony. 2007. Learning to Detect Phishing Emails. *Pages 649–656 of: Proceedings of the 16th international conference on World Wide Web*. ACM.

Frank, Jonas, & Martinez-Vazquez, Jorge. 2014 (January). *Decentralization and Infrastructure: From Gaps to Solutions*. Working Paper 14-05. Andrew Young School of Policy Studies, Georgia State University, International Center for Public Policy, Atlanta, Georgia.

Freund, Yoav, & Schapire, Robert. 1997. A Decision-Theoretic Generalization of On-Line Learningand an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Grimmer, Justin, & Stewart, Brandon M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Guerini, Marco, Strapparava, Carlo, & Stock, Oliviero. 2008. CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing. *Journal of Information Technology & Politics*, **5**(1), 19–32.

Günther, Frauke, & Fritsch, Stefan. 1998. neuralnet: Training of Neural Networks. *The R journal*, 137–142.

Han, Xianpei, Sun, Le, & Zhao, Jun. 2011. Collective Entity Linking In Web Text: A Graph-based Method. *Pages 765–774 of: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM.

Hsieh, Chung-Ho, Lu, Ruey-Hwa, Lee, Nai-Hsin, Chiu, Wen-Ta, Hsu, Min-Huei, & Li, Yu-Chuan Jack. 2011. Novel Solutions for an Old Disease: Diagnosis of Acute Appendicitis with Random Forest, Support Vector Machines, and Artificial Neural Networks. *Surgery*, **149**(1), 87–93.

Jones, Zachary, & Linder, Fridolin. 2015. Exploratory Data Analysis Using Random Forests. *In: Prepared for the 73rd annual MPSA conference*.

Jurafsky, Dan, & Martin, James H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kolodner, Janet L. 1992. An Introduction to Case-Based Reasoning. *Artificial Intelligence Review*, **6**(1), 3–24.

Lantz, Brett. 2013. *Machine Learning with R*. Birmingham, UK: Packt Publishing.

Lautenschlager, Jennifer, Shellman, Steve, & Ward, Michael D. 2015 (March). *ICEWS Coded Event Aggregations*. http://dx.doi.org/10.7910/DVN/28117 Harvard Dataverse Network [Distributor] V1 [Version].

Lautenschlager, Jennifer, Starz, James, & Warfield, Ian. 2016. A Statistical Approach to the Subnational Geolocation of Event Data. *Pages 333–343 of:* Schatz, Sae, & Hoffman, Mark (eds), *Advances in Cross-Cultural Decision Making*. Advances in Intelligent Systems and Computing, vol. 480. Cham, Switzerland: Springer International Publishing.

Li, Jixin. 2003. An Empirical Comparison between SVMs and ANNs for Speech Recognition. *Page 2003 of: The First Instructional Conference on Machine Learning*, vol. 951.

Liaw, Andy, & Wiener, Matthew. 2002. Classification and Regression by randomForest. *R news*, **2**(3), 18–22.

Lucas, Christopher, Nielsen, Richard A., Roberts, Margaret E., Stewart, Brandon M., Storer, Alex, & Tingley, Dustin. 2015. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 254–277.

Maroco, João, Silva, Dina, Rodrigues, Ana, Guerreiro, Manuela, Santana, Isabel, & de Mendonça, Alexandre. 2011. Data mining methods in the prediction of Dementia: A Real-data Comparison of the Accuracy, Sensitivity and Specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests. *BMC Research Notes*, **4**(1), 299.

Mehrotra, Kishan, Mohan, Chilukuri K., & Ranka, Sanjay. 1997. *Elements of Artificial Neural Networks*. Cambridge, MA: MIT Press.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., & Dean, Jeff. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *Pages 3111–3119 of: Advances in Neural Information Processing Systems*Neural Information Processing Systems, for Neural Information Processing Systems.

Minhas, Shahryar, Ulfelder, Jay, & Ward, Michael. 2015. Mining texts to efficiently generate global data on political regime types. *Research & Politics*, **2**(3).

Monroe, Burt L., Colaresi, Michael P., & Quinn, Kevin M. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, **16**(4), 372–403.

Mukherjee, Sayan, Osuna, Edgar, & Girosi, Federico. 1997. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. *Pages 511–520 of: Neural Networks for Signal Processing. Proceedings of the 1997 IEEE Workshop*. IEEE.

Müller, Berndt, & Reinhardt, Joachim. 2012. *Neural Networks: An Introduction*. Berlin: Springer Science & Business Media.

Murphy, Kevin P. 2006. Naive Bayes Classifiers. *University of British Columbia*.

Nadeau, David, & Sekine, Satoshi. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, **30**(1), 3–26.

Ng, Hwee Tou, Goh, Wei Boon, & Low, Kok Leong. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. *Pages 67–73 of: ACM SIGIR Forum*, vol. 31. ACM.

O'Brien, Sean P. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, **12**(1), 87–104.

OEDA. 2016 (September). *Open Event Data Alliance: Phoenix Data Base*.

Porter, Martin F. 1980. An Algorithm for Suffix Stripping. *Program*, **14**(3), 130–137.

Raleigh, Clionadh, Linke, Andrew, Hegre, Håvard, & Karlsen, Joakim. 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset Special Data Feature. *Journal of Peace Research*, **47**(5), 651–660.

Rao, Delip, McNamee, Paul, & Dredze, Mark. 2013. Entity Linking: Finding Extracted Entities In A Knowledge Base. *Pages 93–115 of: Multi-source, multilingual information extraction and summarization*. Springer.

Salehyan, Idean. 2015. Best Practices in the Collection of Conflict Data. *Journal of Peace Research*, **52**(1), 105–109.

Schrodt, Philip A. 2006. Twenty years of the Kansas event data system project. *The political methodologist*, **14**(1), 2–8.

Schrodt, Philip A. 2015 (April). *Event data in forecasting models: where does it come from, what can it do?* Unpublished manuscript.

Schrodt, Philip A., & Yonamine, James E. 2013. A Guide to Event Data: Past, Present, and Future. *All Azimuth*, **2**(2), 5.

Schrodt, Philip A., & Yonamine, Jay. 2012 (September). *Automated Coding of Very Large Scale Political Event Data*.

Shellman, Stephen M. 2008. Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors over Time and Space. *Political Analysis*, **16**(4), 464–477.

T. Joachims, Thorsten. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Pages 137–142 of: European Conference on Machine Learning*. Springer.

Tong, Simon, & Koller, Daphne. 2001. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, **2**(Nov), 45–66.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, Inc.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Vol. 1. New York: Wiley.

Zhang, Harry. 2004. The Optimality of Naive Bayes. *In: Proceedings of the 17th International FLAIRS conference*. American Association for Artificial Intelligence Press.

Zhang, Min-Ling, & Zhou, Zhi-Hua. 2006. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, **18**(10), 1338–1351.

Zurada, Jacek M. 1992. *Introduction to Artificial Neural Systems*. Vol. 8. Boston, MA: PWS Publishing Company.

SOPHIE J. LEE: DEPARTMENT OF POLITICAL SCIENCE
Current address: Duke University
E-mail address: sophie.lee@duke.edu

HOWARD LIU: DEPARTMENT OF POLITICAL SCIENCE
Current address: Duke University
E-mail address: hao.liu@duke.edu

MICHAEL D. WARD: DEPARTMENT OF POLITICAL SCIENCE
Current address: Duke University
E-mail address: michael.d.ward@duke.edu